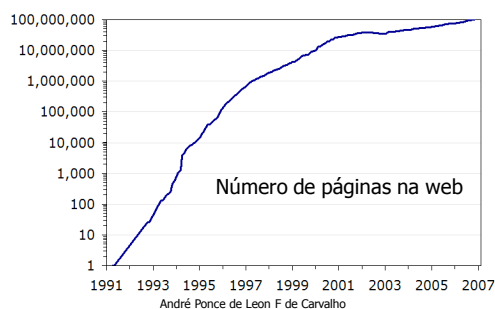
Introdução

- Estima-se que a quantidade de dados em Bases de Dados mundiais dobra a cada 20 meses
  - Transações bancárias
  - Utilização de cartões de crédito
  - Dados governamentais
  - Medições ambientais
  - Dados clínicos
  - Informações disponíveis na web
  - Dados de biologia molecular

André Ponce de Leon F de Carvalho 4

Tamanho Conjunto de Dados

Múltiplos de Bytes		
byte	B	10^0
kilobyte	kB	10^3
megabyte	MB	10^6
gigabyte	GB	10^9
terabyte	TB	10^{12}
petabyte	PB	10^{15}
exabyte	EB	10^{18}
zettabyte	ZB	10^{21}
yottabyte	YB	10^{24}

André Ponce de Leon F de Carvalho 7

Armazenamento de Dados

- Computadores atuais já vêm com 1 ou 2 terabyte de memória
- Cabe em 1 petabyte
  - 20 milhões de arquivos de 4 gavetas cheios
  - 500 bilhões de páginas de texto
    - Metade do conteúdo de todas as bibliotecas acadêmicas americanas combinadas
  - 7 bilhões de fotos no *facebook*
  - 200 milhões de músicas

André Ponce de Leon F de Carvalho 8

Grandes Bancos de Dados

- World Data Centre for Climate (WDCC)
  - Max Planck Institute for Meteorology and German Climate Computing Centre
  - 220 TBs de dados disponíveis na web sobre pesquisas e tendências climáticas
    - 110 TBs (24.500 DVDs) com dados de simulações climáticas
    - 6 PBs de informação adicional em fitas magnéticas

André Ponce de Leon F de Carvalho 9

Grandes Bancos de Dados

- Youtube
  - Pelo menos 45 TBs de vídeos
    - 60% de todos os vídeos assistidos online
  - 100 milhões de vídeos assistidos por dia
  - 65.000 novos vídeos adicionados por dia
    - Crescimento esperado de 1.86 TB por mês

André Ponce de Leon F de Carvalho 10

Grandes Bancos de Dados

- Amazon
  - Dois de seus BDs têm juntos mais de 42 TBs de dados
    - Milhões de itens vendidos por ela e por seus associados por ano
    - Equivale a 37 trilhões de posts para fóruns
  - 59 milhões de consumidores cadastrados

André Ponce de Leon F de Carvalho 11

Grandes Bases Biologia

- GenBank
  - Banco de dados de sequências genéticas do Instituto Nacional de Saúde (NIH)
  - Todas as sequências de DNA publicamente disponíveis
  - International Nucleotide Sequence Database Collaboration
    - DNA DataBank of Japan (DDBJ)
    - European Molecular Biology Laboratory (EMBL)

André Ponce de Leon F de Carvalho 12

GenBank
Crescimento do GenBank
1982-2009
Fonte: Genbank
Seqüências (milhões)
Pares de bases de DNA (bilhões)
Pares de bases
seqüências
1982
1986
1990
1994
1998
2002
2006
André Ponc

GenBank
Crescimento do GenBank
1982-2009
Junho de 2012
>141 bilhões de pares de bases
>154 milhões de sequences
Fonte: Genbank
Seqüências (milhões)
Pares de bases de DNA (bilhões)
Pares de bases
seqüências
1982
1986
1990
1994
1998
2002
2006
André Ponc

Grandes Bases Biologia
Worldwide Protein Data Bank (wwPDB) tem mais de 90.000 estruturas
RCSB PDB (EUA)
PDBe (Europa)
PDBj (Japão)
BMRB (EUA)
2007 7263
2008 7073
2009 7448
2010 7971
2011 8120
No passado, crescimento exponencial
André Ponce de Leon F de Carvalho
15

Introdução
Bases de Dados muito grandes podem conter (esconder) dados preciosos
Existe um interesse crescente em explorar esses dados armazenados
Descobrir conhecimento novo e útil
Apoio a decisão
André Ponce de Leon F de Carvalho
16

Exemplo – Carros
vhigh,vhigh,2,2,small,low,unacc
vhigh,high,3,more,big,low,unacc
vhigh,low,3,4,big,low,unacc
med,low,4,2,small,high,unacc
med,low,3,4,small,med,acc
high,high,2,4,big,med,acc
low,low,5more,4,small,med,acc
low,med,4,4,small,med,acc
low,med,4,4,big,med,good
low,low,4,more,big,med,good
med,low,2,4,small,high,good
low,med,4,4,big,high,vgood
med,med,2,4,big,high,vgood
low,low,5more,more,big,high,vgood
André Ponce de Leon F de Carvalho
17

Exemplo - Carros
Preço
Compra: v-high, high, med, low
Manutenção: v-high, high, med, low
Características técnicas
Conforto
portas: 2, 3, 4, 5-more
pessoas: 2, 4, more
Espaço porta malas: small, med, big
Segurança: low, med, high
Aval. do carro: unacc, acc, good, vgood
André Ponce de Leon F de Carvalho
18

Exemplo - Carros

vhigh,vhigh,2,2,small,low,unacc
vhigh,high,3,more,big,low,unacc
vhigh,low,3,4,big,low,unacc
med,low,4,2,small,high,unacc
med,low,3,4,small,med,acc
high,high,2,4,big,med,acc
low,low,5more,4,small,med,acc
low,med,4,4,small,med,acc
low,med,4,4,big,med,good
low,low,4,more,big,med,good
med,low,2,4,small,high,good
low,med,4,4,big,high,vgood
med,med,2,4,big,high,vgood
low,low,5more,more,big,high,vgood

- Preço
  - Compra: v-high, high, med, low
  - Manutenção: v-high, high, med, low
- Características técnicas
  - Conforto
    - # portas: 2, 3, 4, 5-more
    - # pessoas: 2, 4, more
    - Espaço porta malas: small, med, big
  - Segurança: low, med, high
- Aval. do carro: unacc, acc, good, vgood

André Ponce de Leon F de Carvalho 19

Exemplo - Promotores

+,S10, tactagcaatacgcttgcgttcggtggttaagtatgtataatgcgcgggcttgtcgt
+,AMPC, tgctatcctgacagttgtcacgctgattggtgtcgttacaatctaacgcatcgccaa
+,AROH, gtactagagaactagtgcattagcttatttttttgttatcatgctaaccacccggcg
+,DEOP2, aattgtgatgtgtatcgaagtgtgttgcggagtagatgttagaatactaacaaactc
+,LEU1_TRNA, tcgataattaactattgacgaaaagctgaaaaccactagaatgcgcctccgtggtag
+,MALEFG, aggggcaaggaggatggaaagaggttgccgtataaagaaactagagtccgtttaggt
-, 296, aggcatgtaaacgtcttcgtagcgcatcagtgctttcttactgtgagtacgcaccag
-, 648, ccgagtagacccttagagagcatgtcagcctcgacaacttgcataaatgctttcttg
-, 230, cgctaggactttcttgttgattttccatgcggtgttttgcgcaatgttaatcgcttt
-,1163, tatgaccgaacgagtcaatcagaccgctttgactctggtattactgtgaacattatt
-,1321, agagggtgtactccaagaagaggaagatgaggctagacgtctctgcatggagtatga
-, 663, gagagcatgtcagcctcgacaacttgcataaatgctttcttgtagacgtgccctacg

André Ponce de Leon F de Carvalho 20

Introdução

- Técnicas tradicionais de análise de dados permitem apenas consultas simples
  - Quantos itens de um produto em particular foram vendidos em um dado dia?
  - Não conseguem responder consultas do tipo:
    - Dadas características de um carro, ele é bom?
    - Que tecidos podem estar com tumor?
    - Qual a estrutura terciária de uma nova proteína
  - Técnicas mais sofisticadas, capazes de extrair conhecimento de grandes BD são necessárias

André Ponce de Leon F de Carvalho 21

KDD

- Descoberta de conhecimento em BD
  - Knowledge Discovery in Databases
- Área de pesquisa em expansão
- Teorias e ferramentas computacionais capazes de extrair informação útil de grandes BD
  - Informação útil = conhecimento

André Ponce de Leon F de Carvalho 22

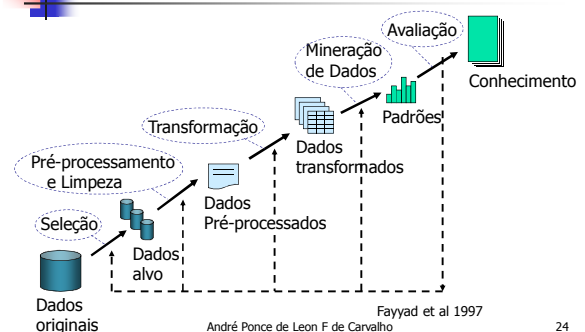
KDD

- Processo de encontrar em dados padrões
  - Úteis
  - Válidos
  - Novos
  - Potencialmente compreensíveis
- Processo interativo e iterativo
  - Várias etapas
    - Uma delas é Mineração de Dados

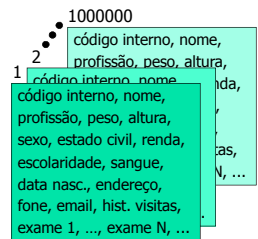
André Ponce de Leon F de Carvalho 23

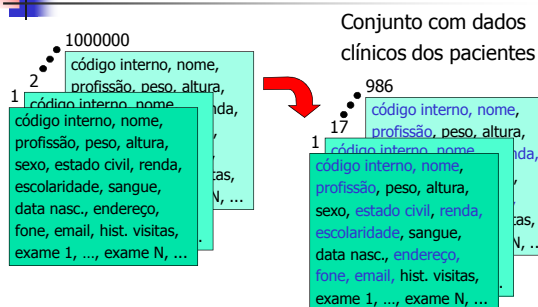
KDD

Avaliação
Mineração de Dados
Conhecimento
Padrões
Transformação
Dados transformados
Pré-processamento e Limpeza
Dados Pré-processados
Seleção
Dados alvo
Dados originais

Fayyad et al 1997

André Ponce de Leon F de Carvalho 24

Seleção
Entender o domínio de aplicação
Determinar o que já é conhecido sobre o problema
Identificar claramente os objetivos do usuário
Exemplo
Diagnosticar um paciente de acordo com um conjunto de sintomas
André Ponce de Leon F de Carvalho
25

Exemplo
BD de um hospital
Composto por conjunto de registros de pacientes
Cada registro é composto de atributos
Informações pessoais
Sintomas
BD com registros de pacientes
1000000
2
1
código interno, nome, profissão, peso, altura, sexo, estado civil, renda, escolaridade, sangue, data nasc., endereço, fone, email, hist. visitas, exame 1, ..., exame N, ...
André Ponce de Leon F de Carvalho
26

Conjunto de Dados
Atributos de entrada (preditivos)
Nome Temp. Idade Peso Altura
João 37 70 94 190 Saudável
Maria 38 65 60 172 Doente
José 39 19 70 185 Doente
Sílvia 38 25 65 160 Saudável
Pedro 37 70 90 168 Doente
Exemplos (objetos, padrões)
Atributo alvo
27

Seleção
Criação de um conjunto de dados
Seleciona "manualmente" um subconjunto dos dados disponíveis
Subconjunto de registros (instâncias ou exemplos)
Subconjunto de atributos considerados relevantes para o problema
Elimina atributos que sejam claramente irrelevantes
André Ponce de Leon F de Carvalho
28

Exemplo
1000000
2
1
código interno, nome, profissão, peso, altura, sexo, estado civil, renda, escolaridade, sangue, data nasc., endereço, fone, email, hist. visitas, exame 1, ..., exame N, ...
Conjunto com dados clínicos dos pacientes
986
17
1
código interno, nome, profissão, peso, altura, sexo, estado civil, renda, escolaridade, sangue, data nasc., endereço, fone, email, hist. visitas, exame 1, ..., exame N, ...
André Ponce de Leon F de Carvalho
29

Pré-processamento e Limpeza
Melhorar a qualidade dos dados e facilitar sua posterior utilização
Engloba várias operações
Seleção "automática" de atributos
Conversão de valores
Lidar com atributos ausentes
Eliminar dados duplicados
Detectar ruído
André Ponce de Leon F de Carvalho
30

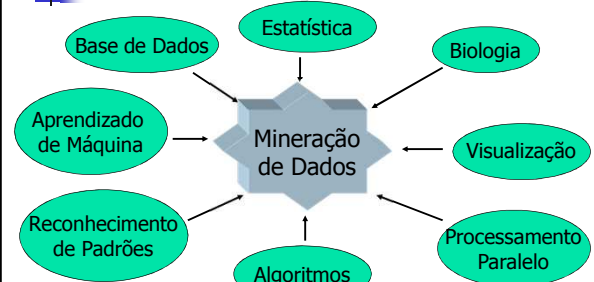
Transformação

- Inclui operações que modificam valores para um dado atributo
  - Cada operação deve ser aplicada a todos os valores do atributo
    - Todos os objetos
  - Ex.: normalização, valor absoluto, ...

Mineração de Dados

- Principal passo no processo de KDD
  - DM e KDD são freqüentemente utilizados como sinônimos
- Fronteiras da etapa de MD no processo de KDD são de difícil identificação
  - Pré-processamento e transformação de dados são freqüentemente vistos como uma parte de MD

Mineração de Dados

- Outros termos utilizados para MD e KDD
  - Extração de conhecimento
  - Descoberta de informação
  - Extração de padrões
  - Análise exploratória de dados

Mineração de Dados

Interpretação / Avaliação

- Interpretação dos padrões minerados na etapa de MD
  - Possível retorno a qualquer uma das etapas anteriores para iteração adicional
- Valida padrões encontrados
  - Importante consulta a um especialista
- Inclui análise estatística
- Ferramentas de visualização têm um papel de suporte importante

Mineração de Dados

- MD X KDD
  - MD: ferramentas básicas utilizadas para extrair padrões de dados
  - KDD: processo que engloba o uso dessas ferramentas, além de:
    - Pré-processamento, seleção e transformação dos dados
    - Interpretação dos padrões
      - Geração de conhecimento

Aplicações

- Número crescente de aplicações
  - Internet: algoritmos de busca, marketing na web
  - Ciência e Medicina: diagnóstico de pacientes, análise de dados do genoma
  - Indústrias: previsão de falhas, diagnóstico de produtos
  - Marketing: segmentação de mercado
  - Telecomunicações: processamento de alarmes, roteamento de linhas de comunicação
  - Finanças: análise de risco, detecção de fraudes, gerenciamento de carteiras de investimento

Aplicações Reais

- Cadeias de PUB britânicas utilizam MD para definir mudanças diárias nos preços de algumas bebidas
  - Acessa impacto das ofertas de *happy hour* nas vendas

Se desconto para uma dada bebida aumenta as vendas em um dia, manter o desconto no dia seguinte, senão tentar outra promoção

Aplicações Reais

- Fast Search & Transfer ASA
  - Software de MD que pode fazer 200 consultas por segundo
  - Utilizado pela Reuters para procurar violações de propriedade intelectual na Web
    - Busca por textos semelhantes aos publicados pela Reuters
    - Envia a Advogados textos suspeitos

Aplicações Reais

- Seimans Medical
  - Ferramenta de MD para o Tratamento de Ataques cardíacos
  - Combina informações médicas de diversas fontes
    - Inclusive texto
  - Busca automática em registros combinados de 6 milhões de pacientes

Aplicações Reais

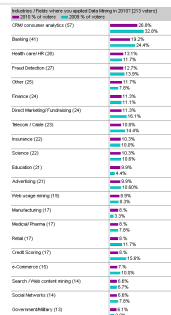
- Seimans Medical
  - Descobriu centenas de casos onde os melhores procedimentos médicos não haviam sido seguidos
    - Mas ainda havia tempo para intervir
  - Identificou pacientes elegíveis para estudos médicos
  - Ganhou o 2005 ICDM Data Mining Practice Prize

Aplicações Reais

- The Mitre Coorporation
  - Ferramenta de MD para detecção de fraudes no imposto de renda
  - Indivíduos com rendimentos elevados são uma das principais fontes de estimativas não realizadas
    - Renda anual > US$ 250.000,00
    - Prejuízo (impostos que deveriam ser coletados – impostos coletados)
  - Sistema de MD usa Aprendizado de Máquina e Análise Estatística para descobrir sonegações

Aplicações Reais
The Mitre Coorporation
Análise baseada em kernels
Utiliza formas conhecidas de burlar a receita
Regras de associação
Procura por grupos de contribuintes que podem estar em um esquema de sonegação
Promovido por um mesmo consultor ou analista financeiro
André Ponce de Leon F de Carvalho
43

Aplicações Reais
The Mitre Coorporation
Modelo trabalha com estimativa de risco, combinando:
Probabilidade de abusos
Potencial de perdas de receita
Resultados
Reduz tempo de análise
2 semanas para poucas horas (dados de 2001)
Encontrou casos não descobertos por auditores
Segundo lugar no 2005 ICDM Data Mining Practice Prize
André Ponce de Leon F de Carvalho
44

Pesquisas KDnuggets
Aplicações de MD
Em que indústrias / áreas você está atualmente aplicando MD em 2010
Fonte:
http://www.kdnuggets.com/polls/2010/analytics-data-mining-industries-applications.html
213 votos
André Ponce de Leon F de Carvalho
45

Investimentos em MD Preditivo
15% - coleta de dados
60% - limpeza de dados
15% - construção e análise de modelos
5% - aplicação
5% - melhorias contínuas
André Ponce de Leon F de Carvalho
46

Produtos de MD
Enterprise Miner
Fair Isaac
Clementine
STATISTICA Version 6
XLMiner
Predictive Dynamix
ANGOSS
WEKA
KnowledgeMiner 5.0
M1
Model 1
KXEN
CART for Windows
wizwhy
DeltaMiner
AnswerTree
PolyAnalyst 4.5
KWIZ
NeuroShell 2
See5
COGNOS
ORACLE 10g
JMP
Data Desk
sas
SPSS
S-PLUS
Resampling Stats
André Ponce de Leon F de Carvalho
47

Mais Produtos
DataEngine
DataMIND
MarketMiner Inc.
pcOLPARS
Urban Science
GainSMARTS
DARWIN
MINESET
PRW
ModelQuest Enterprise
Affinium
The Data Mining Suite
TextSense
Intelligent Miner for data
comotion
DBMiner Insight
NETEZZA
Partek Pro 5.0
48

Mitos (Padhraic Smith)

- "Análise de dados pode ser completamente automatizada"
  - Julgamento humano é crítico na maioria das aplicações
  - Entretanto, semi-automação é muito útil
- "Regras de associação são sempre úteis"
  - Regras de associação são essencialmente listas de correlações
  - Nenhuma aplicação bem sucedida documentada
    - Comparar com árvores de decisão (várias aplicações)

André Ponce de Leon F de Carvalho 49

Mitos (Padhraic Smith)

- "Com uma quantidade massiva de dados, não é necessário estatística"
  - Grande volume leva a heterogeneidade
    - Precisa ainda mais de estatística

André Ponce de Leon F de Carvalho 50

Considerações Finais

- Expansão do volume de dados armazenados
- Necessidade de extrair conhecimento dos dados
- KDD é cada vez mais usado
- Cuidado com promessas exageradas
  - Sistemas Especialistas

André Ponce de Leon F de Carvalho 51

Perguntas

André Ponce de Leon F de Carvalho 52